doi:10.1038/nature09530

Genomic and functional adaptation in surface ocean planktonic prokaryotes

Shibu Yooseph[1]*, Kenneth H. Nealson[1]*, Douglas B. Rusch[1], John P. McCrow[1], Christopher L. Dupont[1], Maria Kim[1], Justin Johnson[1], Robert Montgomery[1], Steve Ferriera[1], Karen Beeson[1], Shannon J. Williamson[1], Andrey Tovchigrechko[1], Andrew E. Allen[1], Lisa A. Zeigler[1], Granger Sutton[1], Eric Eisenstadt[1], Yu-Hui Rogers[1], Robert Friedman[1], Marvin Frazier[1] & J. Craig Venter[1]

The understanding of marine microbial ecology and metabolism has been hampered by the paucity of sequenced reference genomes. To this end, we report the sequencing of 137 diverse marine isolates collected from around the world. We analysed these sequences, along with previously published marine prokaryotic genomes, in the context of marine metagenomic data, to gain insights into the ecology of the surface ocean prokaryotic picoplankton (0.1–3.0 μm size range). The results suggest that the sequenced genomes define two microbial groups: one composed of only a few taxa that are nearly always abundant in picoplanktonic communities, and the other consisting of many microbial taxa that are rarely abundant. The genomic content of the second group suggests that these microbes are capable of slow growth and survival in energy-limited environments, and rapid growth in energy-rich environments. By contrast, the abundant and cosmopolitan picoplanktonic prokaryotes for which there is genomic representation have smaller genomes, are probably capable of only slow growth and seem to be relatively unable to sense or rapidly acclimate to energy-rich conditions. Their genomic features also lead us to propose that one method used to avoid predation by viruses and/or bacterivores is by means of slow growth and the maintenance of low biomass.

Molecular taxonomy and phylogeny[1] revitalized the field of marine microbiology, allowing for the first time the realization that the ‘unseen’ and ‘unknown’ majority of uncultivated microbial taxa could be identified by their 16S ribosomal RNA genes, and identifying widespread clades of marine bacteria and archaea that had no cultivated representatives[2]. More recently, metagenomics has revealed the extent of diversity not fully explained using the available genomes of cultivated marine microbes. When the *Sorcerer II* Global Ocean Sampling (GOS) expedition metagenomic data were published[3,4], it was remarkable in its duality of diversity. On the one hand, there seemed to be only a few taxonomic groups that appeared routinely in marine surface water samples. Of these groups, only three (*Pelagibacter*, *Prochlorococcus* and *Synechococcus*) were represented by cultivated marine microbes. On the other hand, despite their apparent ubiquity and abundance in the surface ocean, there was virtually no complete or near-complete genomic assembly of any of the cosmopolitan taxa, implying that these groups, as judged by 16S rRNA sequence, are internally extremely diverse. The GOS data set was also remarkable because of its apparent lack of relatedness to the entire collection of sequenced genomes: using a process called fragment recruitment, which is akin to *in silico* DNA hybridization, only a few genomes from the entire repertoire of sequenced genomes were found to have a significant number of GOS reads assigned to them.

Here we use a large collection of sequenced marine genomes and metagenomic data to show the presence of two major groups in the marine surface picoplankton with striking differences in their metabolic and physiological capabilities. One group, representing the abundant and cosmopolitan prokaryotic picoplankton, is characterized by small genome sizes and a gene content which suggests that these microbes are capable of only slow growth with little metabolic plasticity. The other group contains a variety of microbial taxa that are found in the surface ocean in low abundance, waiting for nutritionally improved conditions. This group has a gene content that allows for the microbes to adapt to a feast-or-famine lifestyle, and thus occasionally reach high numbers and become a dominant biomass (that is, bloom). We posit that microbes in this group, when considered in an ecological context, are strongly influenced by the presence and/or activities of marine eukaryotes. Our findings also led to a testable hypothesis, which we call ‘cryptic escape’: a major strategy in the true marine picoplankton involves the maintenance of abundant (~10^5 cells ml^{-1}) populations of very small cells (small genomes and low biomass), thus decreasing predation due to bacterivores (ciliates and flagellates) and perhaps bacteriophages.

Results

Overview of the data set and bioinformatics analyses

Our data set (Supplementary Material 1) consists of 197 marine genomes, 10.97 million GOS metagenomic reads and more than 45,000 16S rRNA sequences from GOS 16S PCR libraries and the Ribosomal Database Project[5] (RDP). Of the marine genomes used in this study, 137 were sequenced, assembled and annotated by the J. Craig Venter Institute (JCVI) as part of the Marine Microbial Genome Sequencing Project (MMGSP; http://www.moore.org/microgenome/), which is funded by the Gordon and Betty Moore Foundation. The MMGSP is an international collaborative project that has a primary goal of obtaining genome sequences of ecologically relevant microbes from a variety of diverse marine environments around the world. The MMGSP genomes are a resource for metagenome interpretation and provide insights into the metabolic repertoire and diversity of the marine microbial ecosystem. Many of these microbes were isolated from surface or near-surface waters. The MMGSP genomes thus theoretically complement the metagenomic data gathered by the GOS expedition.

[1] J. Craig Venter Institute, Rockville, Maryland 20850, USA.
*These authors contributed equally to this work.

In the GOS data collection, aquatic microbes were size-fractionated by serial filtration through 20.0-μm, 3.0-μm, 0.8-μm and 0.1-μm filters. The microbes in the 0.1–3.0-μm size range are collectively referred to here as the marine surface picoplankton. The GOS data used in the present analysis are primarily from the 0.1-μm group (92% of the reads); the remainder consists of representation from the 0.8-μm (5% of the reads) and 3.0-μm (3% of the reads) groups. These were collected from the northwest Atlantic/tropical Pacific transect[3] and the later Indian Ocean transect of the *Sorcerer II* GOS expedition.

We assessed the abundance of the sequenced genomes in the picoplankton from their representation in both metagenomic and 16S rRNA PCR libraries that are part of the GOS data set. We used GOS 16S rRNA sequences to assess the relative abundance and geographical distribution of marine taxa that are as yet uncultivated but whose sequencing could explain a larger portion of the GOS data. The 197 sequenced marine genomes were grouped into high- and low-abundance classes on the basis of their occurrence in the GOS metagenomic data. We identified differences in gene content and protein functional groups (pathways and modules) between these classes, and used this to characterize properties associated with the abundant and widespread marine surface picoplankton.

Overlap with GOS data at protein family level

The 137 MMGSP genomes together constitute 552 megabases of DNA sequence and 526,366 proteins (Supplementary Table 1). Various phyla are represented by these genomes, with the most highly represented being Proteobacteria (63.5%), Cyanobacteria (12.4%) and Bacteroidetes (11.6%). These microbes embody a range of physiological diversity and include carbon fixers, photoautotrophs, photoheterotrophs, nitrifiers and methanotrophs. To assess the diversity and representation of proteins in the GOS data, we clustered the 526,366 proteins together with a comprehensive set of proteins from publicly available genomic and metagenomic data sets[6,7]. We found that the MMGSP genomes have high overlap with the GOS data at the protein family level: 78% of the MMGSP proteins fall into clusters that contain GOS sequences and these clusters account for 82.3% of the protein predictions based on GOS reads. On the basis of the clustering, on average 15.4% of the proteins in an MMGSP genome are orfans[8], that is, they do not show sequence similarity to any known proteins. These genomes provide valuable context to inferences made using metagenomic data: 12.4% of previously GOS-only protein families[7] contain MMGSP proteins. We had previously predicted[7] that certain protein domains that were thought to be kingdom specific had examples in other kingdoms; these predictions were verified using the MMGSP genomes. The domains include some previously thought to be specific to eukaryotes (for example the indoleamine 2,3-dioxygenase domain (Pfam ID, PF01231) and the MAM domain (Pfam ID, PF00629)), and a domain previously thought to be specific to archaea (an HTH DNA-binding domain (Pfam ID, PF04967)).

Recruitment of metagenomic reads to sequenced genomes

We used a fragment recruitment tool[3] to assign GOS reads to sequenced genomes, where a recruited read was assigned to a single best-matching genome. Although it is not a phylogeny-based method, fragment recruitment nevertheless helps to assess the representation of a given reference genome or its taxonomic neighbours in a given sample of reads. Of the 10.97 million GOS reads used in this study, 24.5% recruited to the 197 sequenced marine genomes with nucleotide identity matches of ≥50%. The distribution of the recruitment of reads to the genomes is skewed: although the ten most highly recruiting genomes account for 84% of the recruited reads, most of the genomes account for only very small proportions of the recruited reads (Fig. 1). A separate collection of 740 non-marine genomes recruited only 1.5% of the GOS reads.

Recruitment quality was assessed using two measures: depth of coverage, which is defined as the average number of reads covering a base pair in the reference genome, and fraction of the genome covered by the reads. At the 50% identity threshold, 17% of the genomes have a depth of coverage of ≥1 and in 46% of the genomes at least 10% of the genome is covered by the GOS reads (Supplementary Table 2). Only a few organisms in the cultivated set qualify as major constituents in the surface ocean picoplanktonic communities (Supplementary Fig. 1). Most of these are, not surprisingly, in the groups *Pelagibacter*, *Prochlorococcus* and *Synechococcus*, and of these some have very high depths of coverage. A few others are alpha-, beta- or gammaproteobacteria or flavobacteria. A crenarchaeota genome (*Nitrosopumilus maritimus* SCM1) also has a high depth of genome coverage. Overall, however, most cultivated microbes, including those from groups such as *Shewanella* and *Vibrio*, for which there are several sequenced members, are rarely abundant in the surface ocean (Supplementary Material 2).

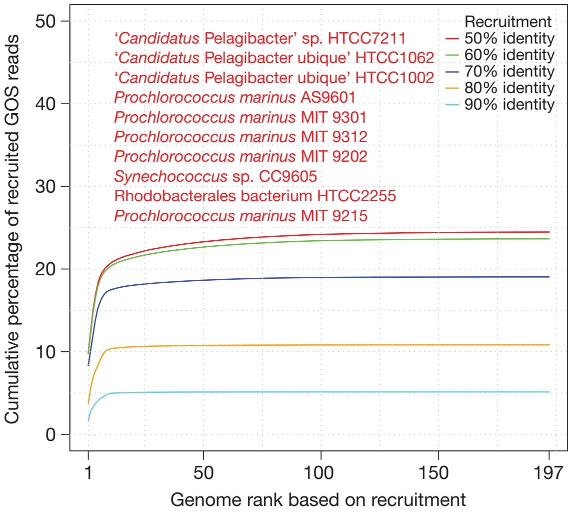

Figure 1 | Fragment recruitment of GOS reads to the 197 sequenced genomes at different nucleotide identity match thresholds. The *x* axis shows genome rank (from highest to lowest) based on the number of GOS reads recruited. The *y* axis shows the cumulative percentage (relative to the total number of GOS reads) of those reads that are recruited. Most of the sequenced genomes are part of the long tail of the distribution. Of the recruited reads, the majority are recruited to a few genomes, with the ten most highly recruiting genomes accounting for 84%, 85%, 91%, 95% and 96% of recruited reads for identity match thresholds of 50%, 60%, 70%, 80% and 90%, respectively. The ten most highly recruiting genomes at the 50% identity match threshold are listed.

Genomes from isolates within the same taxonomic clade recruit metagenomic reads in a pattern consistent with the geography of the sampling locations. The '*Candidatus* Pelagibacter' sp. HTCC7211 strain, which was isolated from the Sargasso Sea, recruits the largest number of GOS reads, in comparison with the cold-water strains HTCC1062 and HTCC1002, which were isolated from the Oregon coast of the Pacific Ocean. The relative recruitment-based ranking of the sequenced isolates changes as we go from the 0.1-μm-filter data to the 3.0-μm-filter data. Whereas the *Pelagibacter* group is dominant in the 0.1-μm-filter data, the *Prochlorococcus* and *Synechococcus* groups become the dominant recruiting groups in the data from the larger filters (Supplementary Material 2). The recruitment patterns for the sequenced isolates across the different GOS sampling sites (Fig. 2) are in agreement with recruitment data[3] from when only a few genomes of each group were available. Of the taxonomic groups that recruit the most, the *Pelagibacter* group is generally present in all GOS samples whereas the *Prochlorococcus* group is notably absent[9] in cold-water GOS samples.

16S rRNA analysis reveals the remaining uncultured majority

Fewer than 25% of the GOS metagenomic reads were recruited to the 197 marine genomes, posing the question of how well the sequenced

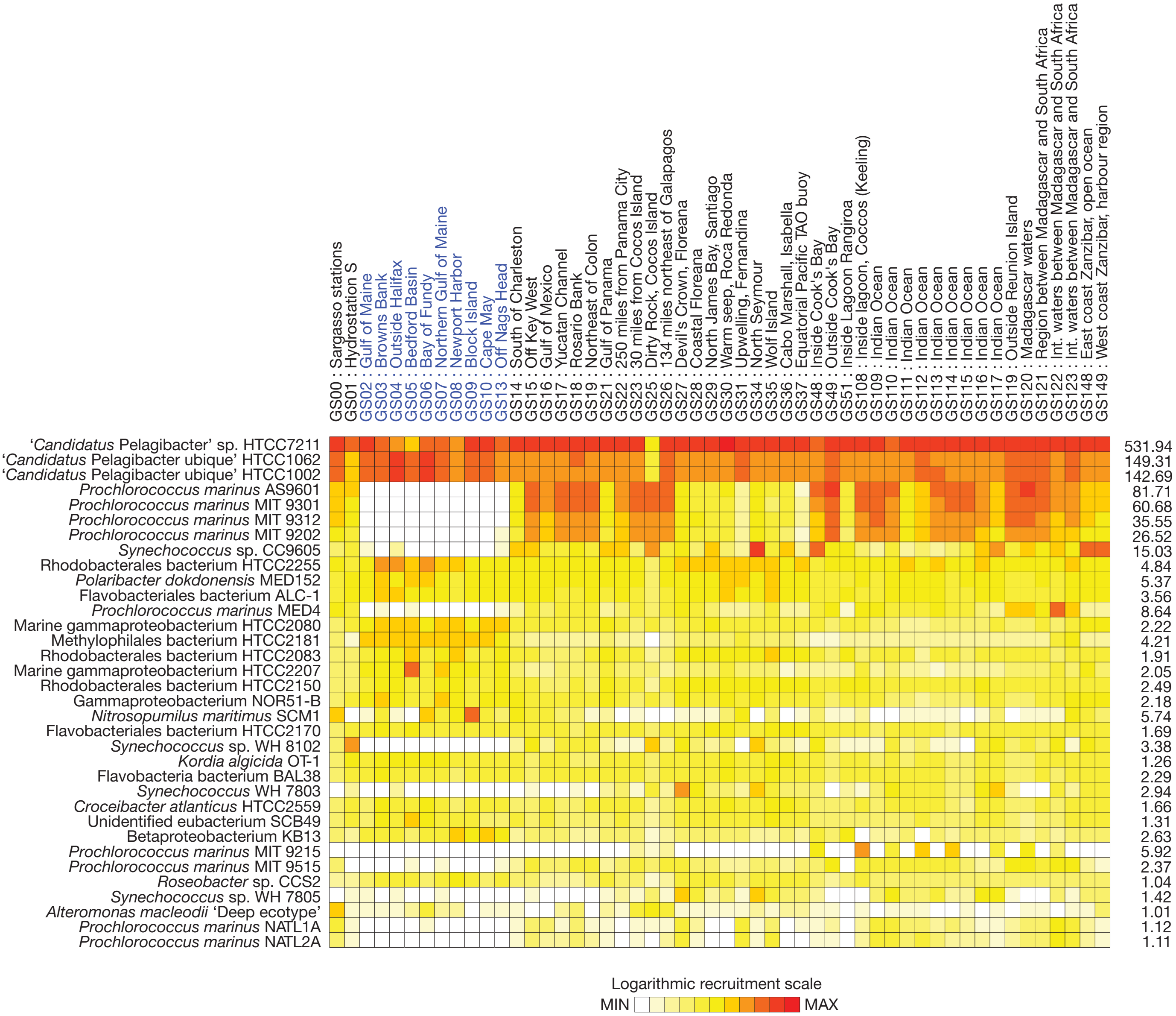

Figure 2 | Abundance and distribution of sequenced marine genomes in different GOS samples, based on fragment recruitment. The raw number of reads from a GOS site recruiting to a genome was normalized assuming 100,000 total reads for each GOS site. Logarithms of these numbers were subsequently taken and each value was assigned a colour relative to the maximum value seen (MAX = 17,623 reads): from light to dark, the colours represent 0 (MIN) to MAX in increments of MAX/10. Normalized values of <100 reads (or 0.01%) are set to 0. The depth of coverage of each genome is shown at the end of its corresponding row. Cold-water GOS samples are highlighted using blue text.

marine genomes represent the 16S-rRNA-based taxonomic space of surface marine bacterioplankton. To determine this, we compared their 16S rRNA genes with 37,860 GOS 16S PCR library sequences plus 8,471 marine bacterial 16S sequences obtained from the RDP[5]. We used a 97% identity match as the cut-off for identification of a bacterial species or operational taxonomic unit (OTU)[10,11]. By this criterion, 15,642 sequences (33.7%) are recruited to 16S sequences of sequenced marine genomes; 16S rRNA genes from 740 non-marine genomes recruit an additional 709 sequences (1.5%). As was observed with fragment recruitment, the recruitment of 16S rRNA sequences from surface marine samples is highly skewed, with most of the 16S sequences recruiting to only a few genomes (Fig. 3).

We determined the phylogenetic distribution of the uncultured OTUs in 16S PCR samples (Methods and Supplementary Material 3). The abundant OTUs belong to bacterial classes that include alpha-, beta-, gamma- and deltaproteobacteria, actinobacteria, flavobacteria, sphingobacteria and cyanobacteria (Supplementary Fig. 2). There are many abundant, but as yet unsequenced, OTUs that are phylogenetically proximal to their sequenced counterparts in the groups *Pelagibacter*, *Prochlorococcus* and *Synechococcus*. The phylogeny also reveals the presence of abundant gammaproteobacteria SAR86[12] (and SAR86-related) and actinobacteria OTUs for which there are no cultured and sequenced representatives as yet. To determine the geographic distribution of the abundant OTUs, we scored the presence or absence of their constituent sequences across 35 GOS sampling sites. Although many of the abundant OTUs are geographically widely distributed in the oceans, several OTUs are cold- or warm-water specific (Fig. 4). The abundant OTUs from this analysis are candidates for isolation and/or sequencing (for instance by single-cell genomics techniques), the better to understand the marine picoplankton and provide more recruitment to the GOS data than we have seen so far. An analysis of 16S sequences from the GOS metagenomic data supports the findings, based on 16S PCR data, on the representation of the sequenced genomes and the abundant uncultured OTUs (Supplementary Material 1 and Supplementary Material 3).

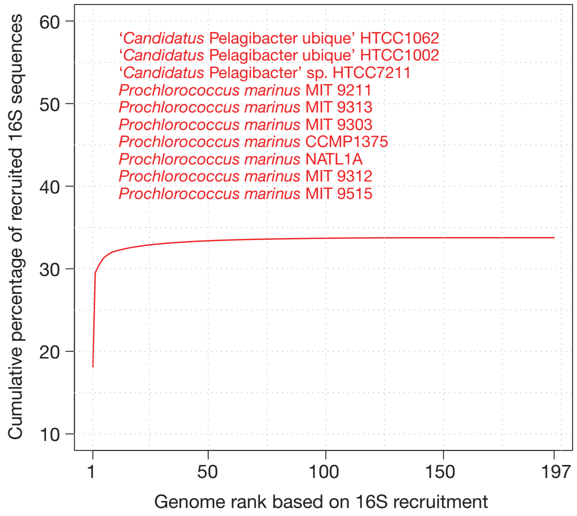

Figure 3 | Recruitment of 16S sequences to the 197 sequenced genomes. The *x* axis shows genome rank based on the number of 16S sequences recruited at the 97% identity match threshold. The *y* axis shows the cumulative fraction (relative to the 48K sequences) of recruited 16S sequences. The ten most highly recruiting genomes at the 97% identity match threshold are listed.

Genomic and functional characterization

To characterize the metabolic and physiological capabilities of the surface picoplankton, we examined the distribution of various genomic features (guanine–cytosine content (%GC) and genome size) and protein families/pathways in the sequenced marine genomes in relation to the depth of coverage of these genomes (Methods, Supplementary Material 4 and Supplementary Fig. 3). We grouped the 197 marine genomes into two sets: the high-recruiting genomes (HRGs), consisting of 34 genomes with coverage $\geq$1, and the low-recruiting genomes (LRGs), consisting of 163 genomes with coverage $<$1. The genomes in the HRG set tend to have smaller size and lower %GC.

We identified protein groups that are differentially distributed between the HRG and LRG sets, even when genome size is taken into account. The slope of the trend line in each category was noted, with a positive slope indicating an overabundance in HRGs for that category and a negative slope indicating an underabundance. All the differentially distributed protein categories (Supplementary Table 3) were organized into higher-order functional groups, which included transcriptional regulation, transport, metabolism, biosynthesis, motility, chemotaxis, secretion, degradation, photosynthesis, and repair and replication.

Our analysis reveals distinct physiological differentiation and offers insights into the functional capabilities of the HRG and LRG sets. One of the most noticeable differences between the two sets is that genomes in the HRG set are characterized by the lack of many functional and regulatory genes. Transcriptional regulation is under-represented, and genes for energy-linked uptake (sugars and amino acids) and efflux (cations, drugs and so on) are nearly absent. Protein secretion to the exterior (for example extracellular proteases, DNases and chitinases) is curtailed, although transport across the cellular membrane to the periplasm seems to be an abundant character. In addition, motility and chemotaxis are absent, and genes known to be involved in quorum sensing (*luxI*, *luxR* and *luxS*), which usually deal with large populations of microbes and/or biofilm populations, are notably absent from the HRG set. Genes coding for functions involved in anaerobic metabolism are greatly decreased in number, whereas genes for aerobic metabolism, intermediary metabolism (glycolysis and gluconeogenesis, the tricarboxylic acid cycle and the pentose phosphate pathway), biosynthesis (amino acids, fatty acids, haem and vitamins), RNA synthesis, and repair and replication seem to be intact.

Transcriptional regulation of nearly every type is greatly decreased in the HRG set, and sensory domains of many types (for example histidine kinases, the GGDEF domain and the methyl-accepting chemotaxis protein domain) are absent or greatly diminished. Also, energy-linked transport systems (for example proton-linked antiporters, PTS-sugar uptake and ATPase-linked efflux systems) are nearly absent from the HRG set, as is transport of amino acids and inorganic ions using ATPase-linked transporters. Similarly, efflux systems for toxic metals and drugs are notably absent.

Another curiosity with regard to transport is the apparent absence from the HRG set of systems for the uptake or synthesis of siderophore-like iron chelates. Because of the small genome size, the genomes in the HRG set have very few iron-requiring proteins and minimal iron requirements[13]. Despite this, iron will still be required for a few fundamental pathways, such as respiration in the heterotrophs, photosynthesis in the autotrophs and the GS–GOGAT cycle in both. On the basis of the genomic characteristics of the HRG set, it seems likely that they take up unchelated Fe(III) or Fe(II). Calculations suggest that this pool should support partial growth of cyanobacteria in even the most iron-depleted oceanic regions[14]. Furthermore, one of the organisms in the HRG set, *Synechococcus* WH8102, may be able to reduce Fe(III)-siderophore complexes, thereby increasing the pool of unchelated iron for uptake[15], although the molecular identity of the reductase is unknown.

The secretion of proteins to the cell exterior is another function that is not well represented in the HRG set. This includes proteins such as serine proteases, metalloproteases, chitinases and DNases (involved in the metabolism of insoluble substrates), as well as pilin and flagellin (self-assembling proteins involved in cell attachment and/or motility). Secretion of proteins to the exterior through the general secretory pathways is virtually absent, suggesting that these microbes are not readily adaptable to the use of insoluble substrates or to a surface-associated lifestyle. However, protein excretion to the periplasm has been not only preserved in the HRG set, but is enriched, suggesting that periplasmic proteins may have an important role in the lives of the marine picoplankters. For example, the twin arginine transport genes (*tat*) and the secretion genes *secA*, *secB* and *secC*, all of which are involved in movement of proteins from the cytoplasm to the periplasm, are well represented in the HRG set.

Perhaps not surprisingly, the genes involved in motility and chemotaxis are strongly under-represented in the HRG set. This includes the structural genes mentioned above (pilin and flagellin synthesis), as well as a large array of sensory genes involved in sensing specific compounds and responding to these compounds through motility and chemotaxis. A number of genes involved in anaerobic metabolism are also notably under-represented in the HRG set. These include genes for low-level oxygen respiration (cytochrome-*bd* complex), genes for nitrate reduction, the anaerobic sigma factor ($^{54}\sigma$), regulatory genes involved in aerobic/anaerobic adaptation (such as *arcBA*) and genes for molybdate uptake and processing.

Clustered regularly interspaced short palindromic repeat (CRISPR) arrays and Cas genes, which form a system involved in bacterial defence against phages, are under-represented in the HRG set, as demonstrated by the counts of CRISPR arrays and genes plotted versus depth of coverage (Supplementary Fig. 3).

Most of the differentially distributed protein categories can be labelled as (−/−), that is, they show negative trends in both the original and the genome-size-normalized data (Supplementary Table 3). Essentially, many protein functions have been lost from the HRG set. However, some protein categories are preferentially retained, as shown by the trends of (−/+) or (+/+); none were (+/−). The positive trends in the normalized data for some categories were the result of a process that is seen in a subset of the genomes in the HRG set (and that is absent from the rest). Examples include protein categories such as carboxysomes and photosystems, which are part of the photosynthesis machinery present in the photoautotrophs in the HRG set. The positive trends seen in other categories, such as the various biosynthesis pathways, reflect core processes that are essential for all of

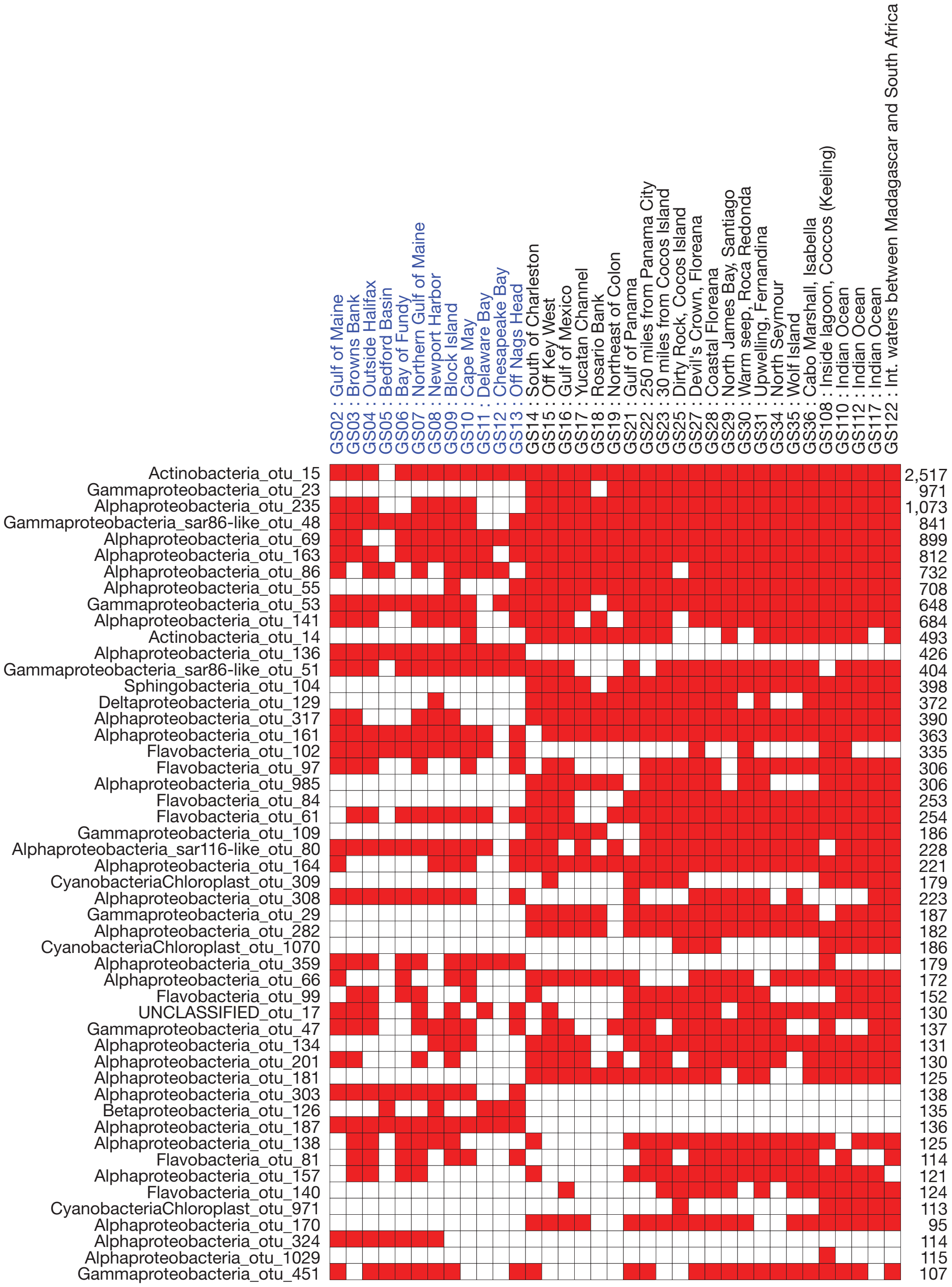

Figure 4 | Distribution of the 50 largest uncultured OTUs in GOS 16S PCR samples. Red and white squares respectively indicate the presence in and absence from a GOS sampling site of the corresponding OTU sequence. The number at the end of each row denotes the number of GOS sequences in that OTU. Cold-water GOS samples are highlighted using blue text.

these genomes. Other positive trends provide insight to the interaction of these organisms with the environment. For instance, photolyases are over-represented in the HRG set, highlighting photodamage as a major environmental pressure in the surface ocean.

There are, as might be expected, a number of categories with strongly positive and/or negative trends that will, as they are identified, be of help in identifying the nature of these HRGs, but many of them are of 'conserved unknown function'. Their identification will be essential in characterizing the microbes and the niches they inhabit.

Discussion

In this Article, we have referred to the 'abundant' and 'cosmopolitan' plankton as the 'true' picoplankton of the surface ocean. Here we distinguish between abundance in a sample and absolute numbers of cells. In many samples in the ocean, organisms with densities of 10^5 cells ml^{-1} or higher are abundant members of the community. Such levels are clearly not high in comparison with microbial densities seen in blooms, which may reach 10^7 cells ml^{-1} or higher: we observe 10^5–10^6 total microbes throughout the oceanic sites, and the 'abundant' microbes represent those taxa that are dominant in this otherwise organism-poor environment. This becomes an important part of the definition of the true picoplankton, as well-known microbial marine genera (for example *Vibrio*, *Alteromonas* and *Photobacterium*), which are surely present in low numbers, were rarely seen in our metagenomic studies. 'Cosmopolitan' implies that a given microbe was very often encountered at the surface sites studied, irrespective of the time of sampling. Because many of the microbes are not seen because of their low abundance, 'abundant and cosmopolitan' as judged by metagenomic data analysis is the descriptive term for what we call the 'true' picoplankton.

The rare and opportunistic biosphere

The work presented here reveals that most of the marine prokaryotic genomes that have been sequenced add little to our appreciation of the biology of the abundant and cosmopolitan picoplanktonic prokaryotes in the surface ocean, other than as a genomic contrast. We regard these low-recruiting microbes as members of a group that is adapted to niches other than the open ocean, including symbiotic, saprophytic, parasitic and other plant and animal associations. Members of this group are predicted to be strongly affected by eukaryotes and/or their products. They have the capacity to survive in the low-nutrient open ocean at low abundance levels (that is, residing in the 'long tail' of diverse organisms seen in the GOS data set), but are able to bloom if presented with the proper energy-rich conditions[16,17]. Such organisms are known to use cell–cell communication mechanisms such as quorum sensing to regulate density-dependent processes such as biofilm formation and other 'group' activities. This notion predicts that the rapidly growing microbes, when found in their natural niches, may be clonal, growing from an original inoculation of one or a few cells. Such a prediction could be routinely tested by examining a number of energy-rich environments in the ocean.

In a recent paper[18], it was proposed that the marine microbial world be divided into oligotrophs and copiotrophs, and that the former "dominate the ocean's free-living microbial populations". Our data support such a view, with the addition that the ability to regulate and adapt to changing conditions should allow the copiotrophs to survive long enough to find another nutrient-rich environment, be it a floating carcass, marine snow, a faecal pellet or the gut tract of a marine eukaryote. Indeed, if every eukaryotic species could harbour a number of species-specific bacterial associates, it would result in an immensely long tail of diversity: microbes awaiting their chance to bloom when conditions improved.

These thoughts are consistent with recent work[19] in which an analysis of the deep-sea bacterial microbial metagenome (at 4,000-m depth near Hawaii) revealed a population more characteristic of the low-recruiting microbes reported here: that is, one rich in genes for motility, secondary metabolism, signal transduction, transport and other high-nutrient-type functions. A simple hypothesis, consistent with these data, is that the deep sea is simply too harsh for the monolithic surface picoplankton, and that the microbes here are again a reflection of a dynamic equilibrium between various high-nutrient environments. Given the sensory mechanisms and motility of the endogenous eukaryotes in the deep sea, they may have a major impact on the aquatic populations of bacteria and archaea observed there. Of particular note would be the issue of whether any of these bacteria and archaea might be of sufficiently low diversity that assembly of their genomes would be possible.

Genome modifications and streamlining

Our data suggest that genome streamlining of the cosmopolitan picoplankton is an important part of their success in the limiting environment of the surface ocean—an observation previously made[20] of members of the SAR11 clade. The few marine microbes that have been characterized suggest that the small genomes have retained virtually all of their biosynthetic abilities and a few key transport systems, but have dispensed with sensory/response systems (chemotaxis, quorum sensing and two-component regulators), motility, anaerobic metabolism and genes involved in organismal interaction. It seems that the marine picoplankton have a very low 'bacterial IQ'[21]. That is, there may be little need for rapid adaptation to changing conditions, with the result that expensive regulatory systems can be dispensed with. It is possible that these microbes, if they regulate their metabolism at all, may engage in regulation using specialized sigma factors, using noncoding RNAs or at the translational (or other post-transcriptional) level in preference to the more expensive transcriptional regulation characteristic of more rapidly growing and adaptive microbes.

This regulatory streamlining, combined with a metabolic single-mindedness, is consistent with the ability to survive as a free-living organism in most of the surface ocean. It seems likely that the uncultured dominant and cosmopolitan picoplanktonic prokaryotes, such as SAR86, have similar overall genomic characteristics but with a different metabolic focus.

It is notable that the approach to genome streamlining seen here is fundamentally different from that seen in other very small microbes, as exemplified by recent work on the genome reduction in *Mycoplasma pneumonia*[22–24]. Although that genome has been streamlined to a very small size, it has occurred without the loss of regulatory mechanisms needed for the bacterium's interactions with 'high-energy' and high-nutrient environments. Many closely related organisms in this group are obligate symbionts or pathogens, which are incapable of host-free growth. Another method of genome streamlining is seen in the recently reported widespread, nitrogen-fixing cyanobacterium UCYN-A[25]. Like the *Mycoplasma* genomes, this small (~1.44-megabase) genome maintains many regulatory and metabolic abilities but discards many of its biosynthetic pathways, suggesting that it may be a symbiont with some as-yet-unidentified metabolic partner[26].

Cryptic escape

We put forward here a hypothesis that we call cryptic escape: success is achieved in the limiting oceanic environment by limiting the effective biomass in such a way as to discourage the success (and evolution) of specific predators, that is, by becoming 'invisible' to them as a food source. This can be done in two ways: by reducing population size and by reducing the amount of biomass per individual. With genome size well correlated with cell size (Supplementary Table 4), a population of SAR11 maintained at a density of 10^5 cells ml^{-1} has a biomass equivalent to less than 10^3 cells ml^{-1} *Vibrio* or *Shewanella*. Whether such a biomass could support the growth of a specific predator is not well documented, and would be an important test of the hypothesis. Cryptic escape thus views one of the driving forces of genome streamlining to be the avoidance of trophic predation, and would be consistent with many different functional (metabolic) end points, a prediction that will be directly tested as the genomes from other (currently uncultivated) cosmopolitan picoplankton groups become available.

It is equally important to know whether the maintenance of low cell numbers and slow growth rates would make viral predation inefficient enough that the CRISPR–Cas system is an unnecessary luxury for the abundant and cosmopolitan picoplankton. At a density of 10^5 cells ml^{-1}, each cell will be thousands of body lengths from others of its species and hundreds of body lengths from the nearest virus (not necessarily a virus specific to that cell), perhaps making efficient viral infection and growth a difficult prospect. With regard to viral predation, many other modes of defence might also be used, such as simply having many fewer metabolic receptors on their surfaces that viruses could use for recognition and entry, and/or extensive microdiversity within the dominant groups, such that annihilation by a single phage becomes nearly impossible. Each of these hypotheses is also testable with the right model systems.

Such an approach does not preclude the existence and success of general predators that can succeed by grazing non-specifically on picoplankters at densities as low as 10^5–10^6 cells ml^{-1}, which are often seen in oceanic waters. Thus, although the carbon fixed and cycled by the picoplankers is of great importance with regard to global carbon cycling, the low levels of biomass may have important consequences with regard to predation (both protistan and viral) of the very small plankters.

However, becoming small and/or rare is not within the purview of the picoplankters. When *Vibrio* (or other heterotrophs) are nutrient limited, they can become very small, often adopting a physiological state referred to as 'viable but not culturable'[27]. Similar states can be seen for many normally fast-growing bacteria when they are nutrient limited at slow growth rates in a chemostat[28]. Thus, it is well within the abilities of the larger heterotrophic bacteria to adopt the cryptic mode, and 'hide' within the long tail of diversity in the ocean.

Conclusion

In essence, the true single-celled picoplankton may be distinguished from most other prokaryotes more appropriately by what they cannot do than by what they can do. That is, the survival strategy seems to be one of dispensing with functions and/or with the control of functions. Thus, a picture emerges of microbes that survive by becoming small, single minded and uncommunicative. Furthermore, the streamlining of genomes is apparently a strategy that is available to many different taxonomic groups of microbes that make up the abundant and cosmopolitan picoplankton. Therefore, it is expected that the other uncultivated cosmopolitan picoplankton groups will show a similar form of genome streamlining, but with metabolic strategies distinct from the other HRGs.

METHODS SUMMARY

MMGSP genome sequencing, assembly and annotation. Two genomic libraries with respective insert sizes of 4 and 40 kilobases were made[29]. The prepared plasmid and fosmid clones were sequenced from both ends on ABI 3730XL DNA sequencers (Applied Biosystems) at the JCVI Joint Technology Center to provide paired-end reads. Successful reads were assembled using the Celera Assembler[30] and the assembly was annotated using the JCVI prokaryotic annotation pipeline (http://www.jcvi.org/cms/research/projects/annotation-service/overview/).

Analysis of functional groups. The depth of coverage at $\geq$50% nucleotide identity threshold (Supplementary Table 2) was used here. For each observed depth of coverage value, c, we binned the genomes into two groups (one containing genomes with coverage $<c$ and the other containing genomes with coverage $\geq c$), and assessed them using the Wilcoxon rank-sum test[31] with the null hypothesis that the given protein category has the same distribution of values in the two groups. This assessment showed that, very frequently, the optimal (smallest) P value corresponded to a coverage of $c = 1$. On this basis, we grouped the 197 marine genomes into two sets: the HRG set, consisting of genomes with $c \geq 1$, and the LRG set, consisting of genomes with $c < 1$. We noted the P values for all protein categories (at $c = 1$), and computed a rejection threshold value (of 4.7×10^{-4}) after correction for multiple testing[32] with the false-discovery rate set to 0.005. We repeated the P-value calculations for all protein categories after normalizing their original values (that is, raw counts) by dividing by the respective genome sizes (as given by the total number of proteins in the genome); here the rejection threshold value was 8.2×10^{-4} for the same false-discovery rate. Subsequently, of the 12,403 protein categories considered, 568 (4.6%) were identified as being differentially distributed in the HRG and LRG sets for both original and normalized data (Supplementary Table 3 and Supplementary Fig. 3), and were analysed further.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 December 2009; accepted 22 September 2010.

1. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).
2. Giovannoni, S. & Rappé, M. in *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 47–84 (Wiley-Liss, Inc., 2000).
3. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5,** e77 (2007).
4. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
5. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37,** D141–D145 (2009).
6. Yooseph, S., Li, W. & Sutton, G. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinform.* **9,** 182 (2008).
7. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
8. Fischer, D. & Eisenberg, D. Finding families for genomic ORFans. *Bioinformatics* **15,** 759–762 (1999).
9. Scanlan, D. J. *et al.* Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.* **73,** 249–299 (2009).
10. Bond, P. L., Hugenholtz, P., Keller, J. & Blackall, L. L. Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. *Appl. Environ. Microbiol.* **61,** 1910–1916 (1995).
11. McCaig, A. E., Glover, L. A. & Prosser, J. I. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl. Environ. Microbiol.* **65,** 1721–1730 (1999).
12. Rappé, M. S., Kemp, P. F. & Giovannoni, S. J. Phylogenetic diversity of marine coastal picoplankton 16S rRNA genes cloned from the continental shelf off Cape Hattera, North Carolina. *Limnol. Oceanogr.* **42,** 811–826 (1997).
13. Dupont, C. L., Yang, S., Palenik, B. & Bourne, P. E. Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proc. Natl Acad. Sci. USA* **103,** 17822–17827 (2006).
14. Morel, F. M. M., Kustka, A. B. & Shaked, Y. The role of unchelated Fe in the iron nutrition of phytoplankton. *Limnol. Oceanogr.* **53,** 400–404 (2008).
15. Lis, H. & Shaked, Y. Probing the bioavailability of organically bound iron: a case study in the Synechococcus-rich waters of the Gulf of Aqaba. *Aquat. Microb. Ecol.* **56,** 241–253 (2009).
16. Nealson, K. H. & Hastings, J. W. Bacterial bioluminescence: its control and ecological significance. *Microbiol. Rev.* **43,** 496–518 (1979).
17. Nealson, K. H. & Venter, J. C. Metagenomics and the global ocean survey: what's in it for us, and why should we care? *ISME J.* **1,** 185–187 (2007).
18. Lauro, F. M. *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc. Natl Acad. Sci. USA* **106,** 15527–15533 (2009).
19. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl. Environ. Microbiol.* **75,** 5345–5355 (2009).
20. Giovannoni, S. J. *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309,** 1242–1245 (2005).
21. Galperin, M. Y. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol.* **5,** 35 (2005).
22. Güell, M. *et al.* Transcriptome complexity in a genome-reduced bacterium. *Science* **326,** 1268–1271 (2009).
23. Kühner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326,** 1235–1240 (2009).
24. Yus, E. *et al.* Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326,** 1263–1268 (2009).
25. Zehr, J. P. *et al.* Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322,** 1110–1112 (2008).
26. Tripp, H. J. *et al.* Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464,** 90–94 (2010).
27. Grimes, D. J., Mills, A. L. & Nealson, K. H. in *Nonculturable Microorganisms in the Environment* (eds Colwell, R. R. & Grimes, D. J.) 209–227 (ASM, 2000).
28. Ingraham, J. L., Maaloe, O. & Neidhardt, F. C. *Growth of the Bacterial Cell* 435 (Sinauer, 1983).
29. Goldberg, S. M. *et al.* A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl Acad. Sci. USA* **103,** 11240–11245 (2006).
30. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
31. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18,** 50–60 (1947).
32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57,** 289–300 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to acknowledge the Gordon and Betty Moore Foundation (grant award agreement #521); the Department of Energy, Office of Science, Office of Biological and Environmental Research (DE-FC02-02ER63453); and the JCVI for funding this research. A.T. and S.Y. also acknowledge support from US National Science Foundation grant 0850256. We also thank the JCVI Software Engineering team for their contribution to the analyses, L. Sheehan for help with manuscript editing, and M. Thiagarajan, L. Foster, I. Singh and M. Lewis for help with data management. Finally, we thank the many collaborators who provided strains for sequencing, including R. Amann, F. Azam, D. Bartlett, R. Belas, C. Bolch, E. Bonch-Osmolovskaya, J. Bowman, I. Brettar, T. Brinkhoff, A. Buchan, R. Case, R. Cavicchioli, P. Chisholm, R. Edwards, D. Emerson, P. Falkowski, C. Francis, B. Fuchs, S. Giovannoni, J. Gonzalez, D. Green, A. Hagström, M. Hart, M. Haygood, R. Hill, M. Höfle, J.-C. Cho, K. Jürgens, S.-J. Kim, G. King, N. King, S. Kjelleberg, M. Klotz, M. Labrenz, E. Lilly, V. T. Marteinsson, M. A. Moran, A. Murray, J. C. Murrell, C. Pedros-Alio, J. Pinhassi, M. Polz, M. Rappé, A.-L. Reysenbach, F. Rodriguez-Valera, E. Rosenberg, N. Ruby, D. Scanlan, H. Schaefer, L. Shimkets, J. Siefert, M. Simon, T. Sokolova, E. Stabb, L. Stal, N. Tandeau de Marsac, B. Tebo, T. Thomas, C. Vetriani, I. Wagner-Dobler, B. Ward, J. Waterbury, B. Wawrik and A. Yayanos.

Author Contributions S.Y., K.H.N., D.B.R., J.P.M., S.J.W., A.T., G.S., A.E.A., C.L.D. and L.A.Z. analysed the data; S.F., K.B. and Y.H.R. sequenced the genomes; M.K., J.J. and R.M. assembled and annotated the genomes; S.Y., K.H.N., E.E., S.J.W., A.T., D.B.R., A.E.A., L.A.Z. and C.L.D. wrote the paper; and J.C.V., M.F. and R.F. were responsible for project management.

Author Information The sequences reported in this work have been deposited in the NCBI Entrez Genome Project database: individual accession numbers can be found in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike Licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.H.N. (knealson@usc.edu).

METHODS

MMGSP genome sequencing, assembly and annotation. Two genomic libraries with respective insert sizes of 4 and 40 kilobases were made[29]. The prepared plasmid and fosmid clones were sequenced from both ends on ABI 3730XL DNA sequencers (Applied Biosystems) at the JCVI Joint Technology Center to provide paired-end reads. Successful reads were assembled using the Celera Assembler[30] and the assembly was annotated using the JCVI prokaryotic annotation pipeline (http://www.jcvi.org/cms/research/projects/annotation-service/overview/).

Fragment recruitment. BLAST[33] was used to recruit GOS reads to sequenced genomes[3]. For the purposes of computing the two recruitment statistics reported here, a recruited read was assigned to a single (best-matching) reference genome.

Alignment and tree building. Separate alignments of bacterial and archaeal sequences were produced using INFERNAL[34]; bacterial and archaeal alignment models from the RDP[5] were used for this purpose. Columns with gaps of >90% were removed from these alignments, and the two alignments were subsequently merged using MUSCLE[35]. This alignment was used to construct a maximum-likelihood phylogeny using RAXML[36].

16S rRNA sequences from the RDP. We identified marine bacterioplankton 16S rRNA sequences from the RDP using a keyword search[37]. Bacterial 16S rRNA sequences ≥1,200 base pairs in length were downloaded and the text information in the GenBank records file was parsed. Only those records that contained word tokens from a positive control set and did not contain words from a negative control set were kept. The positive control set consisted of the following words: marine, coastal, ocean, sea, bacterioplankton, sar11, sar86, sar83, sar116, sar324, sar202. The negative control set consisted of the following words: deepsea, soil, sediment, sand, biofilm, freshwater, pond, lake, hydrothermal, groundwater, borehole, mud, petroleum, marinesnow, aquifer, halophil, oil, diesel, crust, anaerobe, symbiont, hygiene, rhizosphere, associated, viable, biofilter, reactor, sludge, gland, spleen, anoxic, spring, vent, volcanic, basalt, sponge, rock, bog, aquarium, benthic, bone, mat, marsh, mangrove, saltern, urchin.

16S rRNA sequences from GOS PCR libraries. Sequence data were generated using the protocol described in ref. 38. The list of GOS libraries is given in Supplementary Information. Chimaeric sequences were identified and removed. We used two programs to identify chimaeras: a modified version of the RDP chimaera checker[5] and CHIMERASLAYER (http://microbiomeutil.sourceforge.net). Of the sequences that passed chimaera checking, only those ≥1,100 base pairs in length were considered. These sequences were clustered using CD-HIT[39] at high identity and over nearly full length (that is, at ≥99% identity over ≥95% of the length of the shorter sequence). Only sequences in those clusters with five or more members were considered for further analysis; these comprised 37,860 sequences.

OTU identification. The RDP and GOS 16S PCR sequences were searched against 16S sequences from sequenced genomes using BLAST[33]. Those sequences with an identity match of <97% were subsequently clustered using CD-HIT[39]. The clustering was done successively at identities of 99%, 98% and 97%. Each 97%-identity cluster was considered to be an OTU, with size equal to the number of sequences in the flattened cluster. Each CD-HIT representative at 97% identity was considered to be the corresponding OTU's representative.

Determination of the phylogenetic distribution of the uncultured OTUs in 16S PCR samples. We clustered the unrecruited sequences using CD-HIT[39] at 97% identity to produce 1,493 OTUs, with the largest OTU containing 2,748 sequences and 93 OTUs containing ≥50 sequences. Sequences in these OTUs were classified to the class level using the RDP classifier[5]. Those OTUs that contained both RDP and GOS sequences, and those that contained GOS sequences from multiple libraries, were examined further. Representative sequences from these 320 OTUs were aligned, together with 16S sequences of the 197 marine genomes, using INFERNAL[34], and the alignment was used to construct a maximum-likelihood phylogeny using RAXML[36].

Searches against protein databases. The 739,579 proteins from the 197 marine genomes were searched for against several protein databases: COGs[40], Pfams[41], TIGRFAMs[42], KEGG pathways and modules[43], and MEROPS[44]. For searches against COG profiles, a sequence was assigned to its best-matching COG; an E value cut-off of 10^{-8} was used. For Pfam and TIGRFAM assignments, only matches with scores above the model trusted cut-off score were considered. For KEGG assignments, sequences were assigned to KEGG pathways and modules on the basis of best BLAST matches to genes in the KEGG orthologues collection. For MEROPS searches, sequences were assigned to the different families and clans using best BLAST matches to the peptidase units and inhibitor units included in MEROPS. For both sets of BLAST searches, only matches with bit scores of ≥50 were considered.

Analysis of functional groups. The depth of coverage at ≥50% nucleotide identity threshold (Supplementary Table 2) was used here. For each observed depth of coverage value, c, we binned the genomes into two groups (one containing genomes with coverage $<c$ and the other containing genomes with coverage $\geq c$), and assessed using the Wilcoxon rank-sum test[31] with the null hypothesis that the given protein category has the same distribution of values in the two groups. This assessment showed that, very frequently, the optimal (smallest) P value corresponded to a coverage of $c = 1$. On this basis, we grouped the 197 marine genomes into two sets: the HRG set, consisting of genomes with $c \geq 1$, and the LRG set, consisting of genomes with $c < 1$. We noted the P values for all protein categories (at $c = 1$), and computed a rejection threshold value (of 4.7×10^{-4}) after correction for multiple testing[32] with the false-discovery rate set to 0.005. We repeated the P-value calculations for all protein categories after their original values (that is, raw counts) were normalized by dividing by the respective genome sizes (as given by the total number of proteins in the genome); here the rejection threshold value was 8.2×10^{-4} for the same false-discovery rate. Subsequently, of the 12,403 protein categories considered, 568 (4.6%) were identified as being differentially distributed in the HRG and LRG sets for both original and normalized data (Supplementary Table 3 and Supplementary Fig. 3), and were analysed further.

CRISPR–Cas systems. We searched for CRISPR arrays using a modified version of PILERCR (http://www.drive5.com/pilercr/) and our own post-processing scripts for the removal of false-positives using MGTAXA (http://andreyto.github.com/mgtaxa). For the purposes of the current study, we computed several per-genome integral characteristics of the CRISPR system, such as the total number of Cas genes, the total number of CRISPR arrays and the minimal distance between any array and any Cas gene. The source code for the CRISPR analysis pipeline is available as part of our MGTAXA package. Independently of the metagenomic recruitment analysis, we compiled the known lifestyle information for two groups of microbial genomes. The first group consisted of all genomes possessing a significant CRISPR–Cas system (defined as at least three genes and three arrays per genome; $n = 20$), and the second consisted of 13 genomes randomly selected from the total set of genomes lacking CRISPR–Cas features ($n = 130$) (Supplementary Table 5). Within the first group, seven genomes were of deep-sea hydrothermal vent origin, two were from hypersaline environments and eight were associated with particles, surfaces or host organisms, or formed dense colonies in mats or blooms. By contrast, 12 out of 13 genomes lacking CRISPR–Cas were described as free-floating surface picoplankton.

33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
34. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25,** 1335–1337 (2009).
35. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
36. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
37. Hagström, A. *et al.* Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Appl. Environ. Microbiol.* **68,** 3628–3633 (2002).
38. Shaw, A. K. *et al.* It's all relative: ranking the diversity of aquatic bacterial communities. *Environ. Microbiol.* **10,** 2200–2210 (2008).
39. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
40. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
41. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36,** D281–D288 (2008).
42. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31,** 371–373 (2003).
43. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36,** D480–D484 (2008).
44. Rawlings, N. D., Morton, F. R. & Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Res.* **34,** D270–D272 (2006).